Big Data Clustering Model based on Fuzzy Gaussian

Amira M. El-Mandouh
Beni-Suef University
amiramohey@fcis.bsu.edu.eg

Hamdi A. Mahmoud
Beni-Suef University
dr_hamdimahmoud@yahoo.com

Laila A. Abd-Elmegid
Helwan University
drlaila_mohamed@yahoo.com

Mohamed H. Haggag
Helwan University
mohamed.haggag@fci.helwan.edu.eg

***Abstract* — Clustering is also known as data segmentation aims to partitions data set into groups, clusters, according to their similarity. Cluster analysis has been extensively studied in many researches. There are many algorithms for different types of clustering. These classical algorithms can't be applied on big data due to its distinct features. It is a challenge to apply the traditional techniques on large unstructured data. This study proposes a hybrid model to cluster big data using the famous traditional K-means clustering algorithm. The proposed model consists of three phases namely; Mapper phase, Clustering Phase and Reduce phase. The first phase uses map-reduce algorithm to split big data into small datasets. Whereas, the second phase implements the traditional clustering K-means algorithm on each of the spitted small data sets. The last phase is responsible of producing the general clusters output of the complete data set. Two functions, Mode and Fuzzy Gaussian, have been implemented and compared at the last phase to determine the most suitable one. The experimental study used four benchmark big data sets; Covtype, Covtype-2, Poker, and Poker-2. The results proved the efficiency of the proposed model in clustering big data using the traditional K-means algorithm. Also, the experiments show that the Fuzzy Gaussian function produces more accurate results than the traditional Mode function.**

Keywords—Big Data; MapReduce; Fuzzy Gaussian; K-means.

1. INTRODUCTION

Data mining is a mechanism extracting the information from data. It is challenging to get relevant information and provide it within shortage time [4]. In data mining; supervised learning and unsupervised learning are the two learning approaches utilized to mine data [5]. In the Supervised learning; data includes both input and the desired outcome. The desired results are known and are given in inputs to the model during the learning procedure. The neural network, Multilayer perception, Decision tree are examples of supervised models. On the other hand in unsupervised learning. The desired outcome is not given to the model during the learning procedure. This method can be used to cluster the input data in classes by their statistical properties only. These models are for the various type of clustering, k-means, distances and normalization, self-organizing maps.

Data mining had some algorithms like classification, clustering, regression and association rule. Clustering is a task to group data by their similarities and dissimilarities from data elements; mainly it is difficult at the time of big dataset. Clustering method converts that information into various clusters where the object in that group has similar properties as compared to other but not same to other clusters properties.

The rest of this artical is planned as follows. Section II presents related works about the k-means algorithm. Part III contains brief discussion about algorithms which used a new model such as map-reduce, k-means, mod & Gaussian. The proposed algorithm is explained in Section IV. Section V introduces the experimental results by using four big data namely; Covtype, Covtype-2, Poker Hand and Poker Hand-2. In Section VI, we discuss the actual importance of the model in the conclusion section.

2. RELATED WORK

Clustering is a process for partitioning datasets. It is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities [2] [14]. This technique is helpful for an optimum solution. K-mean is the most famous clustering method. Mac Queen in 1967, firstly introduced this algorithm, though the idea went back to Hugo Steinhaus in 1957 [3].

Y. S. Thakare et al. [6] discussed the performance of k-means algorithm which is evaluated with various datasets such as "Iris," "Wine," "Vowel," "Ionosphere" and "Crude oil" dataSets and different distance metrics. It is assumed that performance of k-means clustering depends on the datasets has been used as distance metrics. The k means clustering algorithm is evaluated using recognition rate for a different number of the cluster. This work assisted in choosing suitable distance metric for an appropriate purpose.

SK Ahammad Fahad [7] proposed a method for making the algorithm which is consuming time effective and efficient for reduced complexity. The quality of their resulting clusters heavily depends on the selection of initial centroid and changes in data clusters in the subsequence iterations. After a definite number of iterations, a small part of the data points changes their clusters. Their approach; first gets the initial centroid and sets intervals between those data elements which will not exchange their cluster and those which may exchange their cluster in the subsequence iterations. So, it will decrease significantly in case of large datasets.

Two methods for clustering the large datasets using MapReduce has presented in [8]. Firstly, "K-Means Hadoop MapReduce (KM-HMR)" which focused on the MapReduce implementation of regular K-means. The second one improves the clusters quality to create clusters with distances that are maximum in intra-cluster and minimum in inter-cluster for large datasets. The results of their introduced methodology present enhancement in the execution time efficiency of clustering. Experiments executed on original K-means, and proposed model shows that their approach is both powerful and efficient.

Mugdha et al. [20] introduce an approximate algorithm based on k-means. The algorithm minimizes the complexity measure of k-means by calculating over only those attributes which are of interest is proposed here. Their algorithm cannot manipulate categorical data completely until it is transformed it into equivalent numerical data. Manhattan distance concept has been practiced, which in turn decreases the runtime. It is a new method for big data analysis. Their algorithm is scalable, very fast, and have great accuracy. It succeeded to overcome the disadvantage of k-means of an uncertain number of full iterations. They set a fixed number of iterations, without losing the precision.

G. Venkatesh1 et al.[21] Present a method is called layers three aware traffic clustering based on parallel K-means and the distance metric for minimizing the network traffic cost. Their method applied map-reduce model in three layers. Various algorithms are discussed in their paper; they compared between it, e.g., "Bisecting K-Means", "K-Means Parallel", "Basic K-Means", and "DB-Scan". Their proposed method was done on the same data sets to calculate their execution time and accuracy. It enhances performance by reducing the network traffic using partition, aggregation.

Jerril M. et al. [22] design a proposed algorithm of the parallel K-means algorithm based on map reduce on Hadoop. Their paper compared the performance of evaluation criteria called speedup, scale-up, and Size-up. Speedup tries to evaluate the efficiency of the parallelism to improve the execution time. Scale-up checks the ability of it to grow both the Map-Reduce system and the data size, that is, the scalability of the Map-Reduce tool. Size-up estimates the capacity of it to handle growth. It estimates measurements that take to execute the parallel tasks. According to their opinion, the parallel implementation of K-Means gives better results than sequential K-Means algorithm.

3. Map Reduce paradigm

Map Reduce is the software paradigm for processing larger massive and scalable dataset in the cluster. Map Reduce model processes the unstructured dataset available in a clustering format. Map Reduce is a most popular model used for processing a large set of the data in a parallel and distributed clustering algorithm. It offers numbers of benefits to handle large datasets such as scalability, flexibility and fault tolerance. The map-reduce framework is widely used in processing and managing large data sets. It is also used in such applications like document clustering, access log analysis, and generating search indexes. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two essential tasks, namely Map and Reduce.

A. Mapper Phase

The Map-Reduce framework is commonly used to analyze enormous datasets like tweets sets, online texts or large scale graphs. The Mapper and Reduce are two essential phases in MapReduce algorithm. Firstly, the mapper phase starts the execution of the map-reduce program. The large dataset that passed into a mapping function to create similar small datasets which called chunk [12]. The Mapper uses a list of key/value pairs, then processes all of it. The mapper produces zero or more (key/value) pairs. The mapper phase output contains the key and the value of the number of instances that lied in the dataset. This structure gives a smooth and stable interface for programmers to resolve large-scale clustering difficulties.

Algorithm 1 : Mapper Function

```
Store samples dataset
Do
 Read mapper-data from samples dataset one by one
Do
   clustered data=k-means(samples dataset, K, distance);
   Send clustered data to the Reducer
End Of Mapper-Data
While end-of-file
Call reducer
End
```

B. k-means Clustering Phase

Clustering is considered a core task of exploratory data analysis and applications of data mining. Clustering task is grouping objects' sets in a way that objects in the same group (a cluster) are similar to one another than to those in other groups (clusters) [9] [11]. The Partition clustering is a widely method where a number of objects are set and the data sets are partitioned into a number of clusters in which each cluster includes similar objects.

The k-means algorithm is used extensively for clustering large datasets. The concept is classifying a presented set of data into k number of disjoint clusters, in which the value of k is fixed in advance. The k-means algorithm [6] is effective for many practical applications in producing clusters. However, the traditional k-means algorithm is extremely high in computational complexity, particularly for large sets of data. Moreover, different types of clusters result from this algorithm depending on the random choice of initial centroids. Many attempts were made by researchers to improve the k-means clustering algorithm performance. This paper proposed a method for improving the accuracy and efficiency of the k-means algorithm. It is used widely due to the ability to produce better cluster results compared to other clustering techniques plus its fast computation.

```
Algorithm 2 : K-means
Given: dataset of element (e1, e2… en).K: no clusters,
Target :Split the n data elements into k (≤ n) partitions P = {P1, P2, …, Pk}
1. Set Initial mean value for k cluster randomly.
2. Assign each data element to closest mean.
      $p_i^{(t)}=\left\{e_p:\left\|e_p-m_i^{(t)}\right\|^2\leq\left\|e_p-m_j^{(t)}\right\|^2\forall_{j},1\leq j\leq k\right\},$
3. When data elements have been assigned, the centroid of each of the k clusters becomes the new mean.
      $m_i^{(t+1)}=\frac{1}{\left|p_i^{(t)}\right|}\sum_{zj\in s_i^{(t)}} e_j$
4. Repeat Steps 2 and 3 until when the assignments no longer change.
```

C. Reducer Phase

The reducer phase is the main second part of map reduce. It is responsible for collecting the results coming from mappers. The reducer has three steps; Shuffle, Sort and Reduce.

Shuffle step which receives the output from a mapper phase as input and merges these result tuples into a smaller collection of tuples. In the sort, step values are sorted according to the key. Shuffle and sort process is sent in parallel. The last step here calls the reduce method that takes <key, list of corresponding value> pair and produces the output into the file system.

The reduce phase created a single output. There are multiple reducers to parallelize the aggregations. Finally, MapReduce is considered easier to scale data processing over various computing nodes.

```
Algorithm 3: Reducer function
1. Store clustered data
2. Generate cluster label Vector for clustered data
3. Generate Output Matrix $\boldsymbol{M}\times\boldsymbol{D}$, where M is mappers no. & D is clusters no.
4. Initialize Output label Matrix to all cluster.
5. get output from all MAPPERS
    While hasnext (intermediateValuesIn)
        Put outcome from MAPPERS into output
        Matrix(i)= Output
        Allocate cluster label according to cluster vote
        function Cluster label=Cluster vote(output)
   End While
```

1) Fuzzy Gaussian membership function

The Gaussian fuzzy membership function is considerably famous in the fuzzy logic literature. It considered the main connection between the fuzzy systems and the radial basis function (RBF) of neural networks. Also, the Gaussian is used to represent vague, linguistic terms. It focuses on an adaptive distance measure; it can adapt the distance norm to the underlying distribution of the data which is presented in the different sizes of the clusters [1].Gaussian functions are exercised in statistics to describe the standard distributions. It used in signal processing to represent Gaussian filters. In image processing where two-dimensional Gaussians are performed for Gaussian blurs, in mathematics to solve equations and diffusion equations to define the Weierstrass transform.

```
Algorithm 4 :Fuzzy Gaussian Membership
1. get label Cluster matrix from all mappers
2. Generate a matrix $\boldsymbol{M}\times\boldsymbol{N}$ contains cluster label
   $\boldsymbol{M}$ is cluster no. $\boldsymbol{N}$ is number of mappers
3. Do
    1. Compute "mean" & "standard deviation" for every clustered data
          $\boldsymbol{\mu}=\frac{\sum_{i=1}^{N}X_i}{N}$
          $\boldsymbol{\sigma}^2=\frac{\sum(\boldsymbol{X}_i-\boldsymbol{\mu})^2}{\boldsymbol{N}}$
    2. Compute "Membership function" for every cluster
          $\boldsymbol{\mu}_i(\boldsymbol{x})=\boldsymbol{e}^{-(\frac{X-\mu}{\sigma})}$
          $\boldsymbol{i=1,2,\ldots.M, and\ M\ is\ number\ of\ clusters}$
    3. allocate the cluster label with the greatest membership function
          $\boldsymbol{cluster}=\boldsymbol{Max}(\boldsymbol{\mu}_i(\boldsymbol{x}))$
Until End Of File
```

2) Mode function

It is the majority vote, the concept of mode makes sense for any random variable estimating values from a vector space, containing the real numbers and the integers. The mode-function is quickly comprehensible and accessible to calculate. The clustered label is allocated according to the majority of the clustered data.

$$\boldsymbol{Cluster\ label}=mode\ (output)$$
$$\boldsymbol{mode}=\arg\max[Output]$$

4. PROPOSED MODEL

K-means algorithm is based on determining an initial number of iterations, and iteratively reallocates objects among groups to convergence. The proposed model based on k-means and handled by map-reduce programming model.

The proposed model in this paper consists of two phases as shown in Fig.1 that namely, Mapper Phase and Reduce Phase. The first phase split the big dataset into small groups which called mapper according to RAM capacity. Next, the significant part had started when K-Means received the data from the mapper and return cluster label.

The second phase called reducer phase. In this phase used the Fuzzy Gaussian algorithm and Mode function. It had started after receiving cluster label. So, it collects them to produce one output.

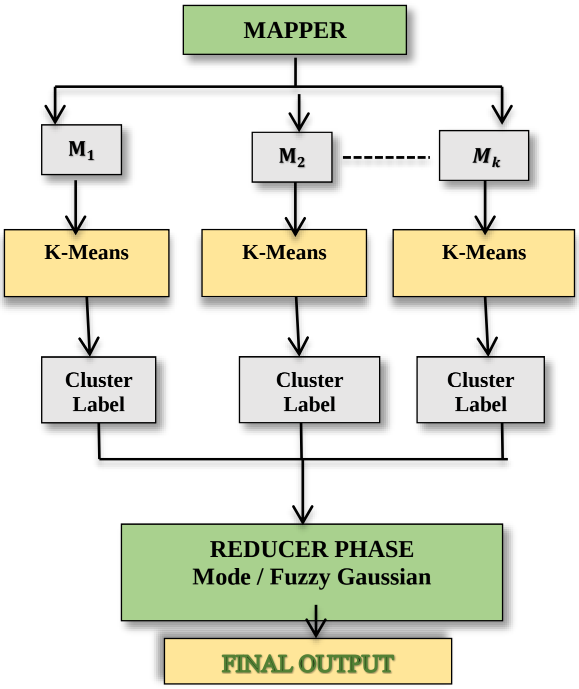

Fig 1 : The flow chart of proposed model.

5. Experimental Results

Data mining algorithms have two important performance indicators are the accuracy for cluster data and the time taken to apply the training.

The propose approach had developed map reduce tool to implement the approach. The experiments are performed on a machine having Intel core i7 processor with 16 GB RAM and windows 10 OS. MATLAB R2014b is used in the experiments.

A. Dataset

In this paper, the datasets had taken from [10]. It is four openly available datasets; Table1 shows the main features of these datasets [10].

TABLE I. Data set Details

DatasetName	Records-no.	Attributes-no.	Classes-no.
"Covtype"	581012	54	7
"Covtype-2"	581012	54	2
"Poker Hand"	1025009	10	9
"Poker Hand-2"	1025009	10	2

The Covtype dataset contains 581012 sample to predict forest cover type from cartographic variables. Any individual relates to one of seven categories (classes) such as "Spruce/Fir, Lodgepole Pine, Ponderosa Pine, and Cottonwood/Willow." The second one is "Covtype-2". It is similar to Covtype except for the number of class (2 class).

Each instance of the Poker-Hand dataset is an illustration of a hand containing five playing cards that drawn from a standard deck of 52. Suit and Rank are two attributes which represent every card, for a total of 10 predictive characteristics. The order of cards is essential, which is why there are 480 possible Royal Flush hands rather than 4. Also, the “Poker Hand-2” is similar Poker Hand except the number of classes is two classes.

B. Results

In this part, experiment's results that are obtained after the implementation of K-means in mapper phase and using two different functions in reducer phase. The four different datasets had applied in the experiments.

TABLE II. Accuracy and time taken by Training the **K Mean and Mode** Algorithm

Data sets / Method	Covtype	Covtype-2	Poker	Poker-2
Accuracy (%)	56.72	62.1	62.67	63.2
Time Taken (Ratio)	7.20126	6.725463	6.012545	6.21541

The results obtained by using Mode function is shown in table 2. The proposed approach achieves 56.72% accuracy in time 7.20126 in case of using "Covtype" dataset which is considered the lowest accuracy and highest time taken. The accuracy improved by 5.38% when decrease the number of classes using "Covtype-2".

TABLE III. Accuracy and time taken by Training the **KM& Fuzzy Gaussian** Algorithm

Data sets / Method	Covtpe	Covtype-2	Poker	Poker-2
Accuracy (%)	62.1	75.6	63.4	73.4
Time Taken (Ratio)	8.01245	8.12542	7.124512	7.124512

The results obtained by using Fuzzy Gaussian are shown in table 3. The best accuracy is 75.6% using "Covtype-2" which enhance results achieved using "Covtype".

Figure 2 shows the comparison between the mode and fuzzy function accuracies have been utilized in reducer phase. The results show improving using fuzzy Gaussian than mod function by leading to simple and straightforward linear algebra implementations. In case if using all " Covtype"," Covtype-2"," Poker "," Poker-2" respectively . The accuracy results indicate that Fuzzy Gaussian is better that Mode function this probably because of allowing one to quantify uncertainty in predictions resulting not just from intrinsic noise in the problem but also the errors in the parameter estimation procedure.

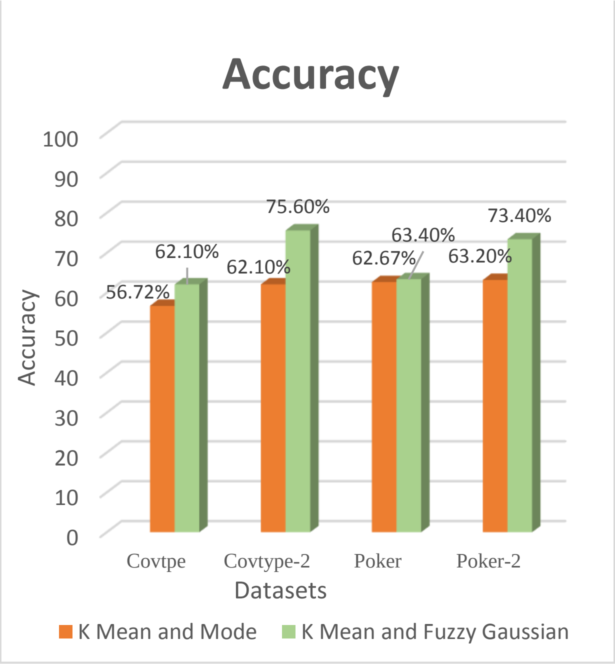

Fig. 2: Accuracy Comparsion for four data set

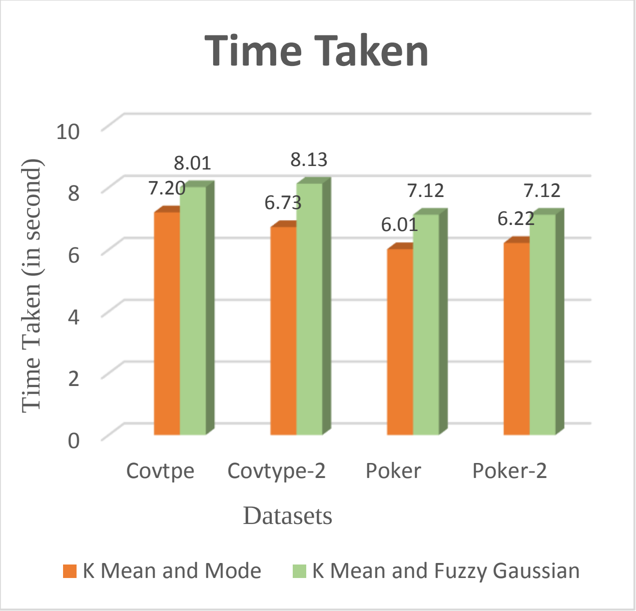

Fig.3: Run Time Comparsion for four data set

Time taken comparisons between the mode and fuzzy function is shown in Figure 3. The results show that the time taken by fuzzy Gaussian is higher than mod function time this probably because of all calculation takes too much time to calculate it more than Mode function.

6. CONCLUSION

The proposed approach is based on the MapReduce programming model. It consist of two phases, Mapper and Reduce phases. In Mapper phase; it had distributed to a mappers group using the map function. K-means is applied on small datasets which existed in mappers. In Reduce phase; the reduce function is resulted by combining outputs using "Mod" and "Fuzzy Gaussian" functions. Gaussian function includes mixed membership; each cluster can have unconstrained covariance structure. Think of rotated or elongated distribution of points in a group. The cluster assignment is flexible. All instance belongs to each cluster to a different degree. The degree is according to the probability of the instance which generated from each cluster's (multivariate) normal distribution. Experimental results showed that the proposed approach gives higher accuracy when using "Fuzzy Gaussian" function than using "Mod" function, as well as perfect time was taken. Also, Fuzzy Gaussian proved its efficiency in accuracy than Mod but with more time in execution.

REFERENCES

[1] Agnes Vathy-Fogarassy, Attila Kiss and Janos Abonyi, "Hybrid Minimal Spanning tree based clustering and mixture of Gaussians based clustering algorithm", pp. 313-330, Springer, 2006.

[2] Neha D., B.M. Vidyavathi, "A Survey on Applications of Data Mining using Clustering Techniques", International Journal of Computer Applications, vol.126, no.2, 2015.

[3] Anshul Yadav, Sakshi Dhingra, "A Review on K-means Clustering Technique", International Journal of Latest Research in Science and Technology, vol.5, Issue 4, no.13-16, 2016.

[4] Vinod S. Bawane, Deepti P. Theng, "Enhancing Map-Reduce Mechanism for Big Data with Density-Based Clustering", International Journal of Innovative Research in Computer and Communication Engineering, vol.3, Issue 4, 2015.

[5] Kosha Kothari, Ompriya Kale, "Survey of Various Clustering Techniques for Big Data in Data Mining", IJIRT, vol.1, Issue 7, 2014.

[6] Y. S. Thakare, S. B. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", International Journal of Computer Applications, vol.110, no. 11, January 2015.

[7] SK Ahammad Fahad, Md. Mahbub Alam, "A Modified K-Means Algorithm for Big Data Clustering", IJCSET, vol.6, Issue 4, 129-132, April 2016.

[8] Chowdam Sreedhar, Nagulapally Kasiviswanath, Pakanti Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data 4, no. 1, 2017.

[9] T. Sajana, C. M. Sheela Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining",Indian Journal of Science and Technology, vol.9, no.3, 2016.

[10] the UCI repository

[11] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering vol.1, 2009.

[12] Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, Francisco Herrera, "Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce", Information Fusion,vol. 42, pp.51-61, 2018.

[13] Mohammed S. Hadi, Ahmed Q. Lawey, Taisir E. H. El-Gorashi and Jaafar M. H. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey", 2018.

[14] Anju, Preeti Gulia, "Clustering in Big Data: A Review", International Journal of Computer Applications, vol.153, no.3, pp.44-47, 2016.

[15] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S. and Bouras, A., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis", IEEE transactions on emerging topics in computing, vol.2, no.3, pp.267-279, 2014.

[16] Yangyang Li, Guoli Yang, Haiyang He, Licheng Jiao, Ronghua Shang, " A study of large-scale data clustering based on fuzzy clustering", Soft Computing, vol.20, no.8, pp.3231-3242, 2016.

[17] Srikanta Kolay, Kumar S. Ray, Abhoy Chand Mondal , "K+ Means : An Enhancement Over K-Means Clustering Algorithm", arXiv preprint arXiv:1706.02949, 2017.

[18] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing, no.1, pp.67, 2016.

[19] Anju Abraham, Shyma Kareem," Security and Clustering Of Big Data in Map Reduce Framework:A Survey", International Journal of Advance Research, Ideas and Innovations in Technology, vol.4, Issue 1, 2018.

[20] Mugdha Jain, Chakradhar Verma, " Adapting k-means for Clustering in Big Data", International Journal of Computer Applications, vol. 101, no.1, 2014.

[21] G. Venkatesh, K. Arunesh, "Map Reduce for big data processing based on traffic aware partition and aggregation", Springer Science and Business Media, 2018.

[22] JERRIL MATHSON MATHEW, JYOTHIS JOSEPH " Parallel Implementation of Clustering Algorithms Using Hadoop", International Journal of Advances in Electronics and Computer Science, vol.3, Issue 6, 2016.